

# Do we really need to catch them all? A new User-guided Social Media Crawling method

Fredrik Erlandsson<sup>1,\*</sup>, Piotr Bródka<sup>2</sup>, Martin Boldt<sup>1</sup>, and Henric Johnson<sup>1</sup>

<sup>1</sup>Blekinge Institute of Technology, Department of Computer Science and Engineering, Sweden

<sup>2</sup>Wrocław University of Science and Technology, Department of Computational Intelligence, Poland

\*fredrik.erlandsson@bth.se

## ABSTRACT

With the growing use of popular social media services like Facebook and Twitter it is hard to collect all content from the networks without access to the core infrastructure or paying for it. Thus, if all content cannot be collected one must consider which data are of most importance. In this work we present a novel User-guided Social Media Crawling method (*USMC*) that is able to collect data from social media, utilizing the wisdom of the crowd to decide the order in which user generated content should be collected, to cover as many user interactions as possible. *USMC* is validated by crawling 160 Facebook public pages, containing 368 million users and 1.3 billion interactions, and it is compared with two other crawling methods. The results show that it is possible to cover approximately 75% of the interactions on a Facebook page by sampling just 20% of its posts, and at the same time reduce the crawling time by 53%. What is more, the social network constructed from the 20% sample has more than 75% of the users and edges compared to the social network created from all posts, and has very similar degree distribution.

## Introduction

We live in an age of big data, in which billions of people are using social media to socialize, interact and create new content at a remarkable rate<sup>1,2</sup>. Online social networks and social media, such as Facebook, Snapchat, Twitter, and Instagram, are used by billions of Internet users. Facebook alone increased its number of users with 12% between 2014 and 2015, and in September 2016 there were 1.79 billion active users on Facebook<sup>2</sup>. This tremendous amount of data is now also available (to some extent) to crawl. However, with limited resources and due to the complexity and speed in which new content is generated, there is a need of better strategies of how and what to collect.

In previous studies we have developed the Social Interaction Network Crawling Engine (*SINCE*)<sup>3,4</sup> that collects publicly available Facebook data. Over a period of two years we have collected 1.6 billion unique Facebook users interacting on 110 million posts through 1.9 billion comments and 31 billion likes. The *SINCE* crawler is novel and unique as it is the first one capable of gathering data in depth by covering all interactions within posts. However, data from social media comes with two inherent problems. The first one, that the data volume is so large that it is close to impossible to continuously gather all content. The second, that only a subset of the data is relevant for a specific application or is interesting to the scientists. In addition, the quality of the data can be evaluated based on different aspects related to the application for which the data is intended to be used, e.g. if it can be used for measuring similarity among users' interactions; if the data provides diversified perspectives on certain topics; or whether the data is a statistically representative sample of the complete data?

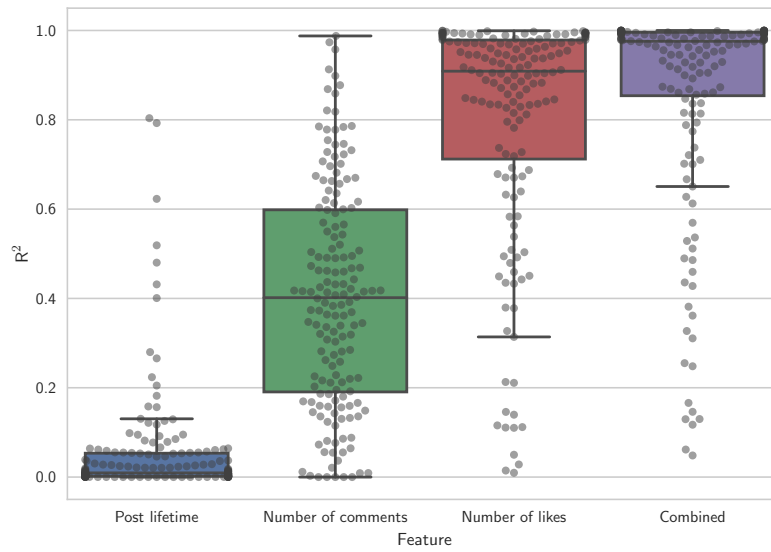
Thus, there is a need depending on the application, to be more strategic when it comes to collect available data (in our case interactions). In this work we define the quality of the data as equivalent with the number of social interactions available, i.e., a user-guided approach is used. Interactions are defined as any user activity (either like, comment or like on comment) towards the post on a Facebook page. Thus, we claim that content quality is a consequence of users' interestingness in that content. There is a lack of research concerning the quality of data in social media and social network research. There are studies on social media and social networks where most of the data is from Twitter. This data is, however, typically collected using Twitter's free garden hose API with a risk of being unbalanced and an unrepresentative sample of the complete data. Studies that address quality of social media data include<sup>5,6</sup>, where the former addresses how social media data from online recommendation systems can be evaluated. Sampling studies of social networks are quite common, including<sup>7,8</sup> that uses the original graph sampling study by Leskovec and Faloutsos<sup>9</sup> as a baseline. Wang et al. presents an interesting study<sup>10</sup> on how to efficiently sample a social network with a limited budget. The study uses metrics of the graph to make informed decisions on how to transverse it. On the topic of graph and social media crawling Zafarani et al.<sup>11</sup> presents ways to evaluate and understand the data generated in social media. However, many social media crawling studies are obsolete due to updates by Facebook, that make it impossible to access a majority of Facebook users' profiles<sup>6,12-14</sup>. More recently Buccafurri et al.<sup>15</sup> discuss different methods to transverse the social network from a crawling perspective by focusing on public groups rather than individual users' profiles. Our approach differs from these studies as we do not create a social network to transverse and only treats the social media as data (not as a graph or network). We focus on data used in studies concerning interactions among users, called Social Interaction Network (*SIN*) graphs<sup>16</sup> as it shows the interactions between users in various communities and can represent interactions of all users on one newsgroup or interacting to a

specific topic.

This study considers publicly available data published in open pages and groups on Facebook. The aim is to investigate how to efficiently and precisely crawl high-quality data from Facebook's social network using the introduced *User-Guided Social Media Crawling (USMC)* method. We investigate if the novel prioritization and ranking techniques in *USMC* can be used to exclude posts that are of low interest during crawling, in order to both reduce the crawling time and at the same time maximise the number of included social interactions. The scope is also to evaluate the proposed *USMC* method to estimate the importance of content, using the wisdom of the crowd and allowing users to point to the crawler which data are of most importance. This way we can try to find a balance between crawling speed and data coverage. Finally, the proposed *USMC* method is evaluated against *random sampling without replacement*<sup>17,18</sup> and a *chronological crawling* method.

## Results

To prioritize data available for crawling, we need to define a set of quality measures which will allow to rank the posts on a page. In this section, we start by testing which of the metadata metrics most accurately assess the importance of a post in terms of how much new knowledge about users interactions on that page it will bring. We continue with using the identified metadata metrics to maximize the number of interactions within the dataset. Finally, we use social networks to validate our findings.



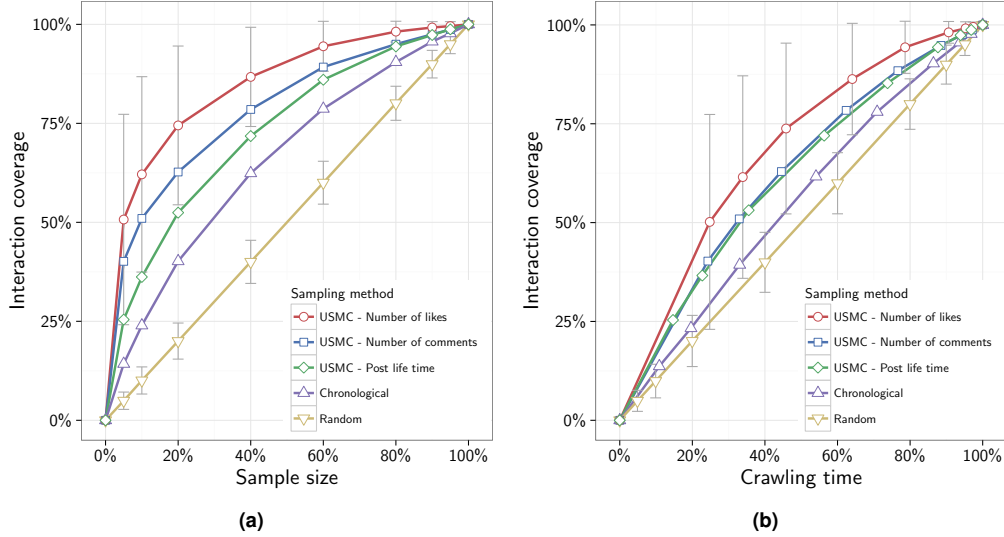
**Figure 1.** This box plot presents the  $R^2$  distribution of OLS Regression test assessing the number of interactions on a *post*, using the following three metadata metrics: *post lifetime* (blue), *number of comments* (green), and *number of likes* (red). The three metrics are also shown as a combined measure (purple). All box-plots are created from a sample of 160 Facebook pages.

The SINCE crawler starts by performing an initial crawl of a page, followed by a full crawl of its data<sup>3,4</sup>. During the initial crawl the SINCE crawler gathers metadata for all posts on a page. For each post the following three metadata metrics are collected: *post lifetime*, *number of comments*, and *number of likes*. An Ordinary Least Square (OLS) regression test<sup>19</sup> was used to investigate which of the three measures most accurately assess the total number of interactions on posts based on the sample of 160 Facebook pages. The basic statistics of this dataset is available in Table 1 and the detailed description of each page is presented in the *Supplementary Information Table S2*.

Figure 1 shows the distribution of  $R^2$  for the OLS regression test conducted, which indicates a high confidence,  $0.80 \pm 0.26$  (std), that the number of likes can be used to predict the number of interactions on a posts. A combination of the three measures produces the most accurate assessment,  $0.86 \pm 0.24$  (std), as illustrated in Fig. 1. However, a combined measure is practically infeasible due to each individual variance per page, which also makes it impossible to use them as a combined global measure. Therefore, in this study we consider the use of each measure alone to assess the number of interactions.

In Fig. 1 there is a clearly visible separation between the distributions of each of the three measures. A Friedman's test  $\chi^2 = 299.73$ ,  $df = 2$ ,  $p \ll 0.001$  shows that there is indeed a statistical difference between the distributions of the three metadata metrics. Further, a Nemenyi *post hoc* test shows (as expected from Fig. 1) that *the number of likes* metric is the strongest predictor for the number of interactions, and that all three distributions are statistically different at significance level 0.001.

As identified in the OLS regression analysis, the *number of likes* is a suitable predictor of the number of interactions on a post. Thus, we use it to rank posts for each page and use that ranking to guide the crawler on which posts it should



**Figure 2.** Average interaction coverage for all 160 Facebook pages for (a) different sample size (represented as a percentage of all post on Facebook a page) and (b) crawling time. It is presented for the *USMC* method with rankings based on *number of likes* (red), *number of comments* (blue), *post lifetime* (green), *chronological crawling* (purple), and *random sampling* method (yellow). For the best and the worst approach we included error bars showing the standard deviation. To see the individual results for all 160 Facebook pages please see *Supplementary Information Fig. S1*.

crawl first. We compare the results in terms of number of collected interactions with a traditional *random sampling without replacement*<sup>17,18</sup> and *chronological crawling* methods. The results presented in Figure 2a shows that by using the *USMC* method it is possible to cover a vast majority of the interactions in a page by considering only a fraction of all posts. For example on average we need to crawl only 20% of all posts to gather 75% of all interactions using the *USMC* method, with posts ranking based on their number of likes. In addition, a sample size of 20% covers only 20% and 40% of the pages' interactions using *random sampling* and *chronological crawling* approaches respectively. For individual results of all 160 Facebook pages please see *Supplementary Information Table S1 and Fig. S1*.

Figure 2b illustrates the fraction of crawling time (x-axis) needed to collect desired portion of interactions. It shows that it is possible to collect more than 50% of the interactions in less than 25% crawling time. That is, approximately twice as many interactions than collected by the *random sampling* and *chronological crawling* methods in the same amount of crawling time. The number of interactions collected in a certain crawling time has a linear relationship for the *random sampling* method. This relationship is more favourable for the *USMC* method as long as the crawling time is less than roughly 80%. For crawling times larger than 80% the gained efficiency over the *random sampling* method decreases fast since the *USMC* method is gathering the least important and meaningful posts from the crawled page.

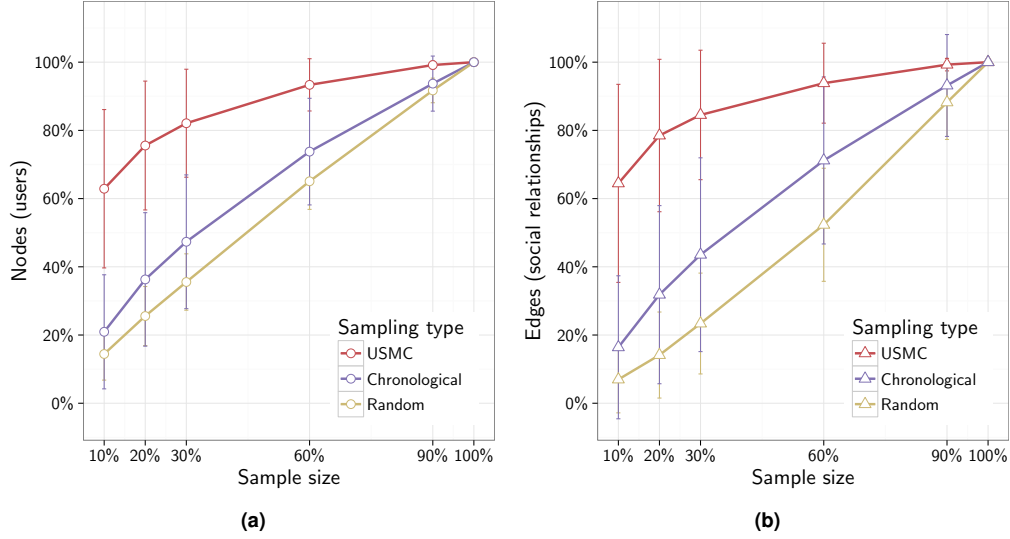
Cohen's *d* values for the findings in Figure 2a show that there are large ( $d > 0.8$ ) differences between the three metadata metrics for all sample sizes smaller than 0.95. Regarding the crawling time, the Cohen's *d* scores show large differences between the metadata metrics for all crawling times shorter than 80%, and medium differences for crawling times longer than 80% but smaller than 95%.

Both Fig. 2a and Fig. 2b show that the most efficient approach is to use *number of likes* metric for ranking posts in the *USMC* method. Thus, the next experiments only consider *number of likes* when comparing the *USMC* method with both *random sampling* and *chronological crawling* approaches.

To further validate the proposed *USMC* method we investigate how complete and useful are the social networks constructed out of gathered data. Please note that due to the computational power limitations we had to exclude the biggest pages from this analysis and completed it for 133 Facebook pages.

We have created three social networks based on the social interactions collected by the *USMC* method as well as the *random sampling* and *chronological crawling* methods for the following sample sizes: 10%, 20%, 30%, 60% and 90% of all posts on a Facebook page. Figure 3 shows the number of *nodes* (a) (Facebook users) and *edges* (b) (interactions between users) in each social network. It is clear that the *USMC* method collects both significantly more users as well as social relations between them, compared to the other two methods. Thus, the social network constructed out of data collected by the *USMC* method is more complete, and even with only 20% of the collected posts we are able to create a network with more than 75% of users and interactions between them.

Next, we have also performed a simple social network analysis with respect to degree distribution for each created network. Fig. 4 presents the degree distribution for three social networks created out of the data from three representative Facebook pages. These three pages are representatives of the first quantile (*Q1*), the Median and the third quantile (*Q3*)



**Figure 3.** The fraction of all *nodes* (a) and *edges* (b) for the social networks constructed based on the 10%, 20%, 30%, 60% and 90% sample of all posts, using *USMC* (red), *chronological crawling* (purple), and *random sampling* (yellow) approaches. The plot is showing mean and standard deviation for all 160 pages.

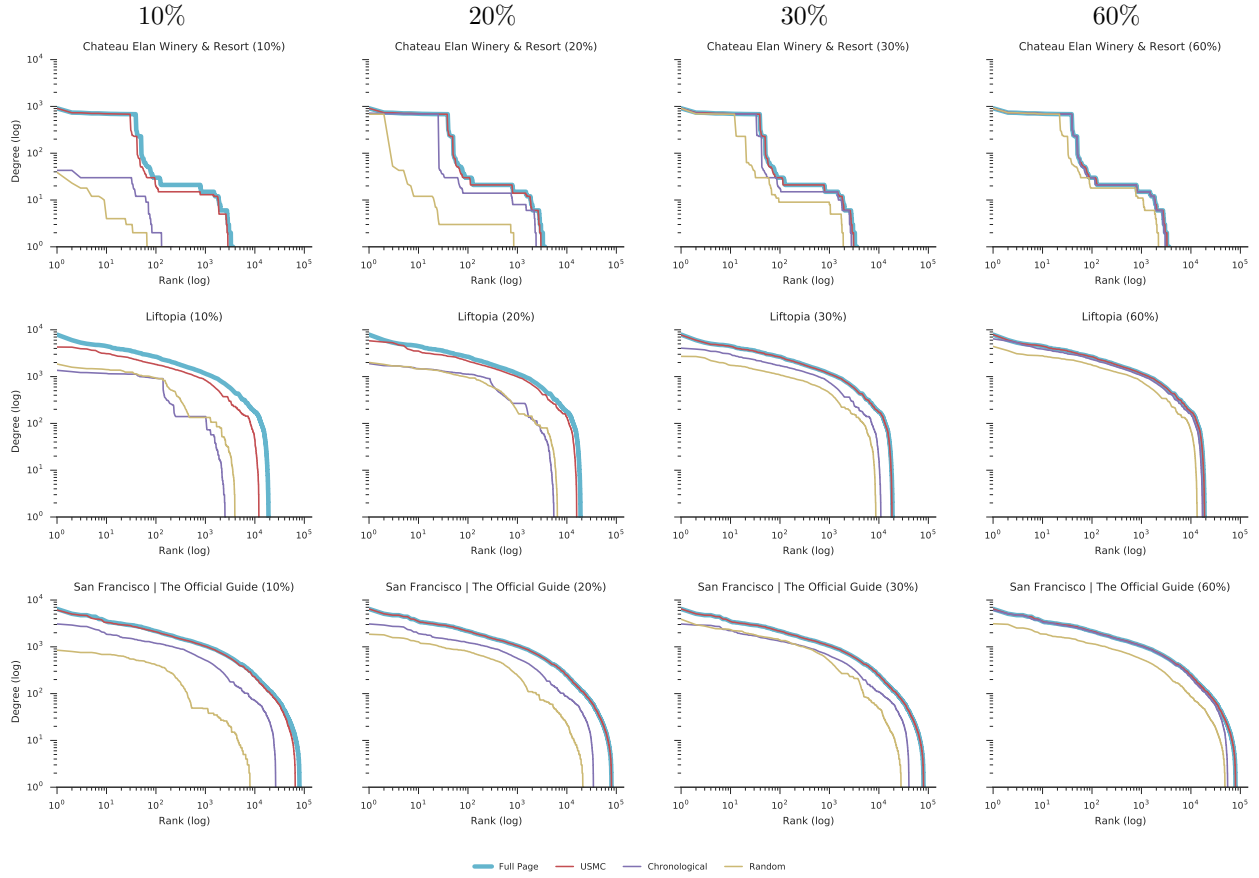
in the number of posts per page distribution for 133 pages. Four different sample sizes, 10%, 20%, 30% and 60% of all posts on a Facebook page, are shown for each of the three pages. Figure 4 shows that even with a relatively small sample of 20% of all posts collected by the *USMC* method it is possible not only to create a social network which not only has more than 75% of all users and interactions between them but, also that the degree distribution is very similar as for the complete social network created from all available data. We can see the same behaviour for all 133 Facebook pages that were analysed, please see the *Supplementary Information Fig. S2*.

## Discussion

In the domain of data gathering it is important to be aware that very often it is not possible to collect all data from large and continuously growing social media sites. Instead one needs to consider when the dataset is “good enough” for once purpose. We have identified two scenarios to consider when deciding on when to stop the data collection. First, *deadline based* data collection when there is a defined time frame, e.g., an upcoming presidential election in four weeks. Following that example, as much data as possible need to be collected as fast as possible (e.g. in two weeks) so, strategic decisions based on the collected on-line behaviour can pinpoint which national regions to focus on before the election day. The second type of data collection strategy is *data coverage* that specifies a particular sample size of the full dataset that is needed, e.g., 75% of the original data is required for credibility of particular study. Thus, if we need 75% of the interactions on a page we could collect the full page which would take 100 days, or use *random sampling* approach and gather data until we reach 75% which will take more or less 75 days, or we can use *USMC* approach and collect the needed 75% of the interactions in 45 days. In that case, we are saving 30 days (Fig. 2b) compared to *random sampling* which is equivalent to a 40% time saving. Of course, other approach is to add more resources to speed up the process but very often is not an available option either due to API restrictions or equipment we have. That is why it is so important to assess which posts need to be collected first by allowing social media users to show which content is of most importance and on what we should focus our limited resources.

The goal of covering the most number of interactions with a limited resources is evaluated in Fig. 2, where we show how big fraction of all interactions can be collected with properly selected sample of all posts. Ranking posts based on *number of comments* covers  $78.5 \pm 16.7\%$  (std) of the all page’s interactions with a sample of 40% of all posts on a page. However, ranking posts based on *number of likes* provides even better coverage, with  $86.7 \pm 12.5\%$  (std) of all interactions at a sample of 40% of all the posts on the page. Figure 2b shows the interaction coverage with respect to crawling time for SINCE crawler. We see that it is possible to decrease the overall crawling time if we can accept not to crawl all posts but only the ones that covers the most number of interactions.

Social media is commonly evaluated using social network analysis<sup>20–22</sup>. With *USMC* method we can cover over 80% of the social network, in terms of nodes and edges, with just 30% of all posts on a page Figure 3. To further evaluate the quality of the network we calculated the degree distribution for each of them. Figure 4s shows three representative pages from our dataset, representing *Q1*, Median, and *Q3*. For all three pages depicted in Fig. 4 the networks produced from posts collected by *USMC* method have close to identical degree distribution already at a 20% sample compared to the whole page. Giving a strong indicator that the social network structure of just 20% of the page’s post are nearly identical



**Figure 4.** The degree distribution for three social networks created out of three representative Facebook pages. Each column shows different sample size (10%, 20%, 30% and 60% of all posts on a page), and each row presents a representative page for each quantile in the number of posts per page distribution for 133 pages. The first row shows the Facebook page *Chateau Elan Winery & Resort* with 1,131 posts, 25,008 users, and 4,814 comments ( $Q_1$ ); the second *Liftopia* with 3,973 posts, 47,001 users, and 50,065 comments (Median); and the third *San Francisco | The Official Guide* with 13,305 posts, 735,183 users, and 116,336 comments ( $Q_3$ ). As we can see with just 20% of all pages collected by *USMC* method, we are able to create a social network which has almost the same degree distribution as social network created from all available data.

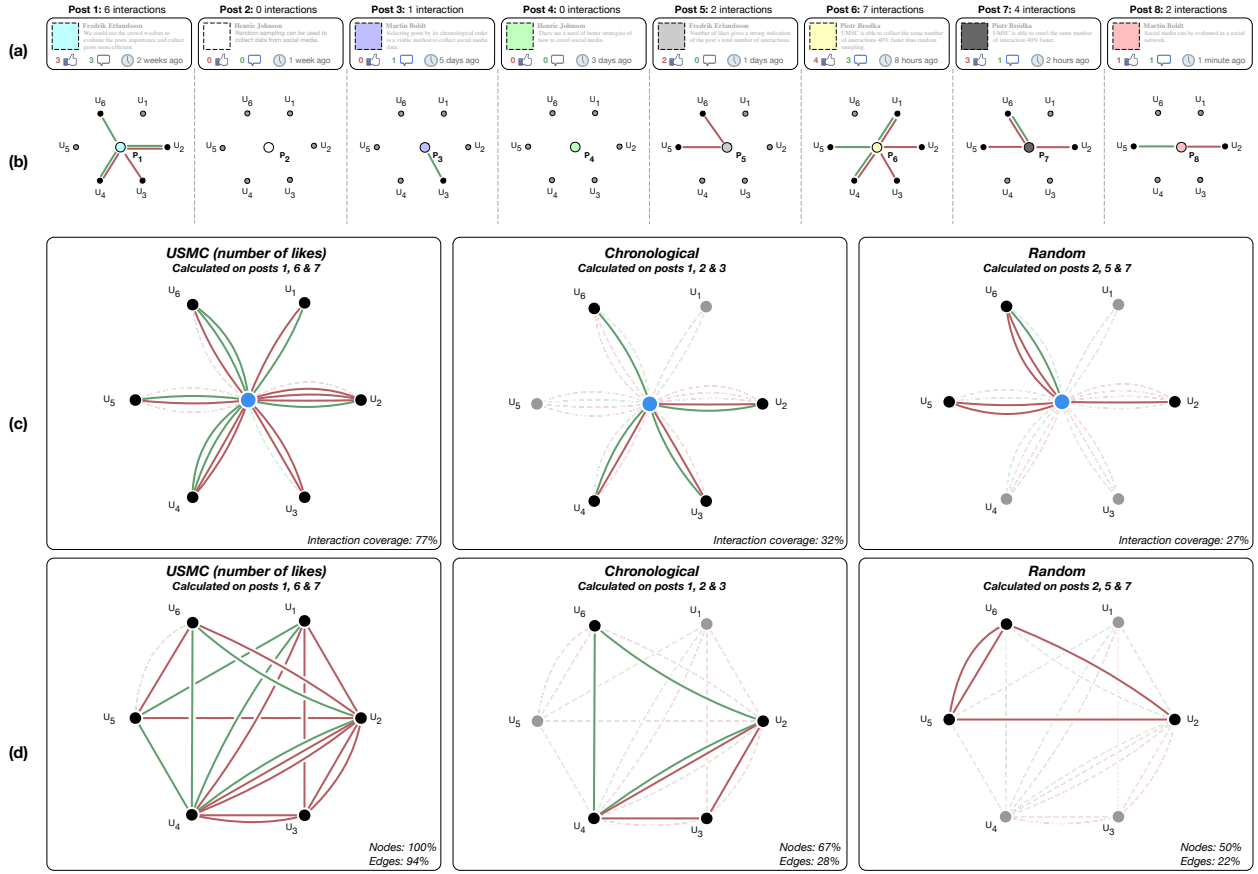
to the complete page. For all pages please see *Supplementary Fig. S2*. Although not every interaction is included in the collected data when using *USMC* approach, it is clear that the resulting social networks are very similar to the network created from all post on a page. That is, both the number of users and the number of social relations are similar, as well as, the network degree distribution. By using *USMC* method it is possible to collect as many social interactions as possible in a limited time, and still get close to identical social network. This is an huge advantage of *USMC*, that is useful when applying for example in deadline-based data collection we have mentioned earlier.

In this study *USMC* method has been evaluated on Facebook's. However, it is possible to apply the same approach on other social media sites, such as Twitter, LinkedIn and ResearchGate. For Twitter, we could rank tweets using the number of re-tweets, likes or responses. With ranking by means of these attributes it is probable that the collection of social interactions from Twitter could be carried out more efficiently and at the same time produce the representative dataset. Similarly, *USMC* could be applied on ResearchGate's social network by ranking content based on the number of comments, RG-score, h-index or average number of downloads per article. Of course, this claims need to be validated with further research on other social media platforms.

## Methods

*User-Guided Social Media Crawling* method (*USMC*) relies on "wisdom of the Facebook users crowd" to find high-quality content in social networks. The introduced crawling technique ranks content in the social network according to what users decide are most interesting for them, i.e. what they choose to interact on. In this work we define social interactions as the type of actions users can take on a posts in Facebook's social network. It can be either a like on a post, a comment on a post or a like on a comment of a post. The social interactions are exemplified in the toy example shown in Figure 5. The eight posts in Figure 5a include different number of interactions with regards to likes (shown as red number next to





**Figure 5.** Example of interactions extracted from posts. Figure (a) shows eight different posts with number of likes ('thumbs-up' icon), number of comments ('speech bubble' icon), and the age of the posts ('watch' icon). Figure (b) shows the bipartite networks of interactions between the six users ( $U_1$ – $U_6$ ) and the eight posts ( $P_1$ – $P_8$ ), where red edges denote likes on posts and green edges denote comments on posts. The users are the same on all posts. Figure (c) shows the aggregated bipartite networks of users interaction towards posts collected by three different crawling methods: *USMC*, *chronological* and *random sampling*. The full network from all eight posts is shown as dashed edges, while the collected interactions are shown as solid lines. Red edges denote likes on posts and green edges denote comments on posts. Figure (d) shows the social networks created based on 37.5% sample of all posts, collected by three different crawling methods: *USMC*, *chronological* and *random sampling*. The full social network from all eight posts is shown as dashed edges, while the collected edges are shown as solid lines.

the 'thumb up' icon) and comments (shown as green number next to 'speech bubble' icon). Figure 5b shows a bipartite network of the interactions between the six users ( $U_1$ – $U_6$ ) and each respective post, where green edges represent comments from users on a particular post while red edges represent likes.

As we have users interacting on social media we can use the actions from users (likes and comments) to rank posts. Evaluating data from social media can be made in various forms, but it is hard to computationally evaluate the content. Hence we are using the users' actions to make informed decisions of the social media data in *USMC*, i.e. benefiting from wisdom of the crowd. Users actions on posts gives an indicator of how interesting the post is for the users in the particular community (different communities can have different values and understandings of the subject). *USMC* is using this information to rank which posts to include and in which order.

The dataset used for evaluating the *USMC* method was created by collecting 160 randomly selected open pages on Facebook. Table 1 shows descriptive statistics for these 160 pages included in the study. We used the Social Interaction Network Crawling Engine (SINCE)<sup>3,4</sup> for collecting the 160 Facebook pages between July 2014 and May 2016. SINCE is designed to collect publicly open pages using Facebook's API. The resulting dataset has a median page size of 5,235 posts, 180,314 users, 45,592 comments and 442,424 likes. In total, the dataset includes some 368 million unique users interacting in little over 1.3 billion social interactions. 25 out of the 160 pages were too large to be represented as social networks in memory given the hardware constraints of 148 GB of RAM we had. Additionally, two pages were too small since there must be at least 10 users interacting on two or more posts. Thus, we removed those 27 pages from the social network evaluation part. For complete statistics of all pages please see supplementary information Table S2.

We evaluate the *USMC* method by comparing it to both traditional *random sampling without replacement* and *chronological* methods. *Random sampling* is collecting posts at random, this gives traditionally the best representation of

the data (sampled data will represent the original dataset given the current sample size). We run each *random sampling* execution 100 times and report the mean and standard deviation. The *chronological* method simply collects the posts from the oldest to the newest. Looking at the conceptual example in Figure 5 and having time to collect only 37.5% of all posts, i.e., 3 out of the 8 posts, the *USMC* (ranked by number of likes on posts) will collect the three posts with highest number of likes, i.e. post 1, 6 and 7, while the *chronological* method will collect post 1, 2 and 3 and the *random sampling* will collect post 2, 5 and 7.

**Table 1.** Descriptive statistics of the dataset of 160 pages.

Metric	Mean	Std.	Min	Q1	Median	Q3	Max	Sum
Posts	21,590	60,786	7	1,313	5,235	14,758	470,528	3,454,456
Users	2,588,338	11,460,281	182	26,589	180,314	897,564	113,379,978	414,134,086*
Comments	608,873	2,584,414	37	10,638	45,592	230,205	27,550,352	97,419,710
Likes	7,640,858	33,972,674	384	54,923	442,424	2,589,165	308,495,988	1,222,537,425

\* Of which are 368,094,952 unique users, not overlapping on different pages.

Each page is evaluated with regards to the number of interactions those methods cover. Five different sample sizes (10%, 20%, 30%, 60% and 90% of all post at the Facebook page) are used to represent the page. In the evaluation we also investigate the time it takes to crawl the 160 Facebook pages. In our toy example in Figure 5, each method produces a different post set: *USMC* 1, 6, 7, *chronological* 1, 2, 3 and *random sampling* 2, 5, 7. These three different post sets contain different number of interactions Figure 5c. Posts 1, 6 and 7 collected by *USMC* contain 77% of all interactions while, the posts 1, 2 and 3 (*chronological*) contain only 32% and posts 2, 5 and 7 (*random sampling*) only 27% of all interactions.

Furthermore, we evaluate the three methods by creating and comparing social networks based on the collected interactions from each technique. In these social networks, users are represented as nodes and the edges between them represent social interaction. The social network is created as a undirected graph as  $G = \langle \mathcal{V}, \mathcal{E} \rangle$ , with a set of nodes  $\mathcal{V} = \{v_1, \dots, v_n\}$  to represent users and a set of edges  $\mathcal{E} = \{ \langle v_i, v_j \rangle : v_i, v_j \in \mathcal{V} \wedge i \neq j \}$  representing relationship between the users  $i$  and  $j$ . An edge is present if both users  $i$  and  $j$  are commenting the same post.

Figure 5d shows the resulting social networks created by three crawling methods (*USMC*, *chronological* and *random*) using the same sample size of 37.5% of all posts, i.e., of 3 out of the 8 posts. It is clear that *USMC* creates the most complete network since it includes all of the six existing users and 94% of edges. The *chronological* and *random* methods include 67% and 50% of the nodes (users), and 28% and 22% of the edges respectively. Please note that in our toy example for presentation purposes we have shown multilayer social network with two types of edges based on (i) likes and (ii) comments represented in form of multigraph Figure 5d. However, as we have mentioned earlier, in our experiments the social network is a singlelayer network where edges are based on comments only.

Finally, a summary of the statistical tests used for evaluation purposes are as follows. First, we used an ordinary least square regression test<sup>17</sup> to investigate which metadata metrics (out of *post lifetime*, *number of comments*, and *number of likes*) is most accurate in predicting the number of interactions on posts. Secondly, the non-parametric Friedman test<sup>18</sup> is used to identify overall differences in the data since it is not normally distributed. Thirdly, a Nemenyi post-hoc test<sup>18</sup> is used to identify individual differences between metadata metrics. Finally, all reporting of results includes standard measurements such as the test statistic, p-value, mean/median and standard deviation.

## References

1. Twitter.com. Company | about. URL <https://about.twitter.com/company/>.
2. Facebook.com. Company info | facebook newsroom. URL <http://newsroom.fb.com/company-info/>.
3. Erlandsson, F., Nia, R., Boldt, M., Johnson, H. & Wu, S. F. Crawling online social networks. In *Network Intelligence Conference (ENIC), 2015 European*, 9 – 16 (IEEE, 2015). DOI 10.1109/ENIC.2015.10.
4. Erlandsson, F. & Wu, F. S. Socialcrawler. <https://github.com/dslfaithdev/SocialCrawler> (2016). URL <https://doi.org/10.5281/zenodo.153825>. DOI 10.5281/zenodo.153825.
5. Agichtein, E., Castillo, C., Donato, D., Gionis, A. & Mishne, G. *Finding high-quality content in social media* (ACM, New York, New York, USA, 2008).
6. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P. & Bhattacherjee, B. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, 29–42 (ACM, 2007). DOI 10.1145/1298306.1298311.

7. Gjoka, M., Kurant, M., Butts, C. T. & Markopoulou, A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *IEEE INFOCOM 2010 - IEEE Conference on Computer Communications*, 1–9 (IEEE, 2010). DOI 10.1109/INFOCOM.2010.5462078.
8. Gjoka, M., Butts, C. T., Kurant, M. & Markopoulou, A. Multigraph Sampling of Online Social Networks. *IEEE J. Sel. Areas Commun. (USA)* **29**, 1893–1905 (2011). DOI 10.1109/JSAC.2011.111012.
9. Leskovec, J. & Faloutsos, C. Sampling from large graphs. In *the 12th ACM SIGKDD international conference*, 631–636 (ACM, New York, New York, USA, 2006). DOI 10.1145/1150402.1150479.
10. Wang, X., Ma, R. T. B., Xu, Y. & Li, Z. Sampling online social networks via heterogeneous statistics. In *IEEE INFOCOM 2015 - IEEE Conference on Computer Communications*, 2587–2595 (IEEE, 2015). DOI 10.1109/INFOCOM.2015.7218649.
11. Zafarani, R., Abbasi, M. A. & Liu, H. *Social Media Mining. An Introduction* (Cambridge University Press, 2014).
12. Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. Crawling Facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 52:1–52:8 (ACM, New York, NY, USA, 2011). DOI 10.1145/1988688.1988749.
13. Wilson, C., Sala, A., Puttaswamy, K. P. N. & Zhao, B. Y. Beyond Social Graphs: User Interactions in Online Social Networks and their Implications. *ACM Transactions on the Web (TWEB)* **6**, 17–31 (2012). DOI 10.1145/2382616.2382620.
14. Crnovrsanin, T., Muelder, C. W., Faris, R., Felmlee, D. & Ma, K.-L. Visualization techniques for categorical analysis of social networks with multiple edge sets. *Social Networks* **37**, 56–64 (2014). DOI 10.1016/j.socnet.2013.12.002.
15. Buccafurri, F., Lax, G., Nocera, A. & Ursino, D. Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences* **256**, 126–137 (2014). DOI 10.1016/j.ins.2013.08.046.
16. Nia, R. *et al.* Sin: A platform to make interactions in social networks accessible. In *Social Informatics (SocialInformatics), 2012 International Conference on*, 205–214 (IEEE, 2012). DOI 10.1109/SocialInformatics.2012.29.
17. Walpole, R., Myers, R., Sharon, M. & Ye, K. *Probability & Statistics - For Engineers and Scientists*. Pearson (Cambridge University Press, 2012).
18. Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures* (Chapman & Hall/CRC, 2011), 5 edn.
19. Davidson, R. & MacKinnon, J. G. *Econometric Theory and Methods* (Oxford University Press, USA, 2004).
20. Safko, L. *The social media bible: tactics, tools, and strategies for business success* (John Wiley & Sons, 2012).
21. Kietzmann, J. H., Hermkens, K., McCarthy, I. P. & Silvestre, B. S. Social media? get serious! understanding the functional building blocks of social media. *Business horizons* **54**, 241–251 (2011). DOI 10.1016/j.bushor.2011.01.005.
22. Saganowski, S. *et al.* Predicting community evolution in social networks. *Entropy* **17**, 30 – 53 (2015). DOI 10.3390/e17053053.

## Acknowledgement

This work was partially supported by the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316097 [ENGINE], by The Polish National Science Centre, the decision no. DEC-2016/21/D/ST6/02408 and by the Faculty of Computer Science and Management, Wrocław University of Science and Technology statutory funds.

## Author contributions

FE created an initial concept of user-guided social media crawling; FE and PB developed the concept to its current state; FE and PB designed the experiments; FE implemented and executed all experiments and simulations; FE, PB and MB analysed data and discussed results; All authors drafted, critically reviewed the manuscript and approved the final version.